IOSVLM: A 3D Vision-Language Model for Unified Dental Diagnosis from Intraoral Scans

Huimin Xiong[1†], Zijie Meng[1†], Tianxiang Hu[1], Chenyi Zhou[1], Yang Feng[2], and Zuozhu Liu[1✉]

[1] Zhejiang University, Hangzhou, China
[2] Angelalign Technology Inc., Shanghai, China
{huimin.21,zuozhuliu}@intl.zju.edu.cn

Abstract. 3D intraoral scans (IOS) are increasingly adopted in routine dentistry due to abundant geometric evidence, and unified multi-disease diagnosis is desirable for clinical documentation and communication. While recent works introduce dental vision-language models (VLMs) to enable unified diagnosis and report generation on 2D images or multi-view images rendered from IOS, they do not fully leverage native 3D geometry. Such work is necessary and also challenging, due to: (i) heterogeneous scan forms and the complex IOS topology, (ii) multi-disease co-occurrence with class imbalance and fine-grained morphological ambiguity, (iii) limited paired 3D IOS–text data. Thus, we present IOSVLM, an end-to-end 3D VLM that represents scans as point clouds and follows a 3D encoder-projector-LLM design for unified diagnosis and generative visual question-answering (VQA), together with IOSVQA, a large-scale multi-source IOS diagnosis VQA dataset comprising 19,002 cases and 249,055 VQA pairs over 23 oral diseases and heterogeneous scan types. To address the distribution gap between color-free IOS data and color-dependent 3D pretraining, we propose a geometry-to-chromatic proxy that stabilizes fine-grained geometric perception and cross-modal alignment. A two-stage curriculum training strategy further enhances robustness. IOSVLM consistently outperforms strong baselines, achieving gains of at least +9.58% macro accuracy and +1.46% macro F1, indicating the effectiveness of direct 3D geometry modeling for IOS-based diagnosis.

Keywords: IOS · Unified diagnosis · VLM.

1 Introduction

Intraoral scans (IOS) is rapidly becoming routine in clinical dentistry [5,12]. Its high-fidelity 3D surface geometry preserves fine-grained tooth–gingiva morphology beyond conventional 2D imaging, enabling more consistent and auditable assessment of subtle abnormalities [10]. Clinically, a single scan often contains multiple co-existing diseases, requiring integrated multi-disease diagnosis and

† Equal contribution. ✉ Corresponding author.

natural-language reporting for documentation and dentist–patient communication—beyond segmentation or single-lesion detection [19].

Recent dentistry work explored vision-language models (VLM) for unified diagnosis. DentVLM [13], DentalGPT[3], and OralGPT [25] support visual question-answering (VQA) and generative reporting on diverse 2D dental images, providing a general interface for fusing evidence into natural-language conclusions. However, they do not directly model 3D surface geometry, where many abnormalities present as fine-grained morphological changes. OralGPT-Omni[8] and ArchMap [24] instead render 3D scans into multi-view images and applies 2D VLMs, but this relies on view selection and can weaken spatial relationships and geometric cues. Hence, a key gap remains: **end-to-end multi-disease diagnosis with language generation from native 3D IOS inputs.**

However, direct 3D IOS–based unified diagnosis faces challenges. First, inputs are highly heterogeneous: single-arch and occluded-arches scans acquired practically differ substantially in coverage and occlusion/contact visibility, and IOSs exhibit complex morphology and topology, complicating representation learning [6,10,22]. Second, multiple diseases often co-exist in one scan. Class imbalance and subtle geometric differences exacerbate inter-disease confusion, demanding unified reasoning from local morphology to global semantics [13]. Third, 3D IOS-text paired data and high-quality annotations is rare [11,15]. Existing public resources are insufficient for systematic 3D VLM training and evaluation.

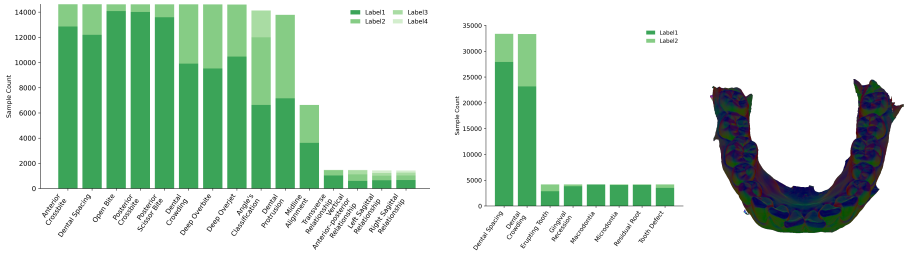
To this end, we propose IOSVLM, a 3D VLM model that directly consumes native 3D IOS geometry for unified multi-disease diagnosis and generative VQA, together with IOSVQA as a paired training and benchmark suite that covers 23 oral diseases and 2 common clinical settings, single-arch and occluded-arches scans, totalling 19,002 IOS cases and 249,055 VQA pairs. To our knowledge, it is among the most comprehensive IOS diagnostic VQA resources to date, explicitly capturing the real-world setting of co-existing diseases and heterogeneous inputs.

IOSVLM adopts a 3D encoder–projector–LLM framework where the 3D encoder captures multi-scale geometric features, the projectors map them into the LLM token space, and the LLM generates communicable diagnostic outputs. Importantly, IOS often lacks reliable color and common 3D pretraining implicitly assumes colored point clouds, naively dropping or padding color causes distribution shift and weaken cross-modal alignment. We thus introduce the geometry-to-chromatic proxy to compensate for that, to improve fine-grained morphology encoding and language alignment. IOSVLM is trained with a curriculum-style schedule: pre-training on larger but noisy supervision to build 3D perception and geometry–language alignment, and fine-tuning on higher-quality supervision to enhance reliability and interpretability under realistic label noise and limited explanations. Experiments show that IOSVLM substantially outperforms strong state-of-the-art baselines, improving macro accuracy by at least 9.58% and macro F1 by at least 1.46%. Our main contributions are:

1. We construct the first large-scale multi-source IOS diagnostic VQA dataset that supports multiple scan type inputs for multi-disease diagnosis.

Table 1. Statistics of IOSVQA and its source datasets. S: single-arch. O: occluded-arch. SD: single-arch disease. OD: occluded-arch disease. # denotes the case number.

Dataset	Privacy	Source	IOS types	#SD	#OD	#cases	#VQA
MaloccIOS	Private	China	S+O	2	11	14,630	208,370
DiseaseIOS	Private	China	S	8	—	4,172	33,376
Bits2Bites	Public	Italy	O	—	5	200	7,309
IOSVQA	—	Mixed	S+O	8	15	19,002	249,055



(a) Occluded-arches (b) Single-arch (c) GCP

Fig. 1. Disease statistics of occluded-arches IOSs (a) and single-arch (b) on IOSVQA dataset, and visualization of normal-based geometry-to-chromatic proxy (GCP) (c). Labels 1-4 denote disease-specific annotation categories used only for visualizing the within-disease class distribution; the same label index may correspond to different clinical meanings across diseases.

2. We introduce the first end-to-end VLM that takes native 3D IOS geometry as input for unified diagnosis, achieving clear performance advantages.
3. Geometry-to-chromatic proxy is proposed to bridge the distribution gap between color-free IOSs and color-dependent point-cloud pretraining, improving fine-grained geometry encoding and cross-modal alignment.

2 Method

2.1 Problem Formulation

IOSVLM targets two common IOS inputs: single-arch scans and occluded-arches scans (refer to Fig. 3). Let $\mathcal{D} = \{1, \ldots, C\}$ denote the set of oral diseases, where $C = 23$. Each disease $d \in \mathcal{D}$ corresponds to a multi-class label set Y_d. Given an input pair $(\widetilde{M}, q_d)$ with IOS mesh $\widetilde{M}$ and question q_d, the diagnosis is formulated as a generative VQA task that produces an answer $A_d \in \{y, (y, r)\}$ $(y \in Y_d)$, i.e., a label y or a label–rationale pair (y, r). Since multiple diseases may co-exist within a single scan, each scan is paired with multiple disease questions. Training and evaluation are conducted over scan–disease pairs as disease-level instances.

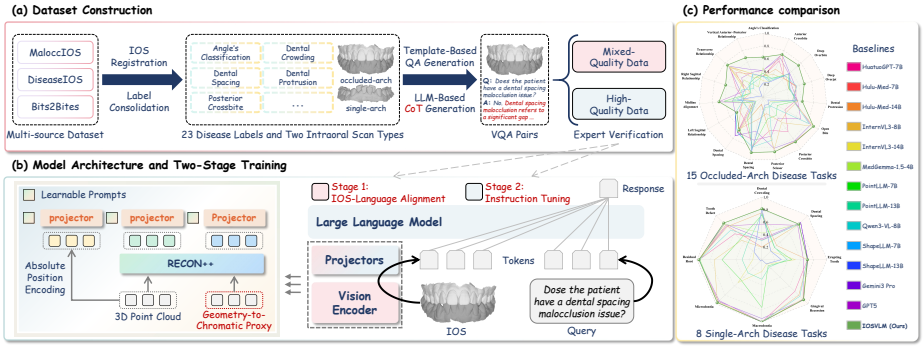


Fig. 2. Overview of our framework and dataset construction.

2.2 IOSVQA Dataset Construction

We first construct a large-scale VQA dataset IOSVQA. It comprises 19,002 IOS cases and 249,055 QA pairs spanning 23 disease categories, with disease statistics detailed in Fig. 1. Data were aggregated from three sources: MaloccIOS, DiseaseIOS, and Bits2Bites [2] (Table 1). To make geometric representations comparable across sources, IOSs were globally registered to standardize orientation and relative maxillomandibular pose. Each case provides one IOS paired with a disease label to form QA samples; questions were sampled randomly from a predefined list and answers correspond predominantly to disease labels.

Disease Label Consolidation. For MaloccIOS, 2D imaging diagnoses were extracted from clinical reports. A randomly selected subset of 557 cases was manually corrected by 28 orthodontists, yielding 7,628 high-quality samples, while the remaining 200,742 samples contain partial label noise. Senior dentists defined rule-based mappings to convert 2D diagnoses into IOS labels. Labels for DiseaseIOS and Bits2Bites were annotated by 5 and 1 orthodontic experts, respectively, and are treated as high quality.

Data Split and Rationale Strategy. IOSVQA is divided into Stage-1 training, Stage-2 training, and testing, with each stage covering all three data sources to maintain disease diversity. The 7,628 high-quality MaloccIOS samples are fully assigned to Stage-2. For interpretability, approximately 50% of Stage-2 samples are augmented with GPT-4o–generated chain-of-thought (CoT) rationales. Motivated by prior evidence that limited CoT supervision can induce reasoning behavior [23], we apply rationale annotation to a subset of high-quality data to balance cost, efficiency, and reasoning performance. To mitigate scale imbalance, the smaller Bits2Bites dataset is augmented by pairing each case with 9 and 3 questions for Stage-1 and Stage-2 construction, respectively.

2.3 Model Architecture

We adopt the multi-modal VLM architecture that combines a large-scale pre-trained 3D encoder with multi-branch projectors and LLM (Fig. 2). The 3D

encoder injects strong geometric priors, while the structured projectors map multi-granularity visual features into semantic tokens aligned with the LLM. The LLM then performs reasoning to produce diagnostic outputs, enabling unified perception and explainable diagnosis across heterogeneous IOS inputs.

To map the raw IOS mesh $\widetilde{M}$ into the semantic space of LLM, we first convert it to a point cloud $\widetilde{P}$ by taking the gravity center point of each face. $\widetilde{P}$ is then randomly down-sampled to obtain P with N points. A pre-trained 3D point-cloud encoder ReCon++ [14] extracts complementary absolute position embeddings F_{ape}, local geometric features F_{local}, and a global descriptor F_{global}. They are mapped into the LLM token space via three dedicated MLP projectors $\phi_{\text{ape}}, \phi_{\text{local}}, \phi_{\text{global}}$, and each projected feature is concatenated with a corresponding learnable visual prompt $V_{\text{ape}}, V_{\text{local}}, V_{\text{global}}$ to balance their contributions:

$$F_p = \left[V_{\text{ape}}; \phi_{\text{ape}}\left(F_{\text{ape}}\right); V_{\text{local}}; \phi_{\text{local}}\left(F_{\text{local}}\right); V_{\text{global}}; \phi_{\text{global}}\left(F_{\text{global}}\right)\right]. \quad (1)$$

The fused point tokens F_p are concatenated with text tokens and jointly fed into the LLM, enabling unified language–visual representations.

Geometry-to-Chromatic Proxy. Point-cloud related models are often pretrained with xyz + RGB inputs. In practical IOS pipelines, downstream storage and exchange typically preserve geometry while discarding or de-emphasizing color/texture, making appearance cues unavailable. Removing RGB thus introduces a pretraining mismatch and discards channels the encoder has learned to exploit. Moreover, RGB is often beneficial not due to semantic color, but as a local separability cue that supports boundary discovery and stable local matching. We therefore propose *Geometry-to-Chromatic Proxy (GCP)*: a geometry-derived proxy that mimics this separability component, enabling better reuse of color-pretraining priors while remaining purely geometry-driven.

We instantiate GCP using surface normals, which encode local surface orientation and capture curvature changes etc. For each point i with normal n_i, we define a flip-robust mapping $\text{GCP}(n_i) = \left|\frac{n_i}{\|n_i\|_2}\right| \in \mathbb{R}^3$. The ℓ_2 normalization standardizes magnitude, and the absolute value removes sign ambiguity, avoiding spurious color inversions. Mapping $\text{GCP}(n_i)$ onto the IOS as pseudo-colors (Fig. 1(c)) yields spatially coherent patterns with clear transitions at anatomical boundaries, supporting that GCP provides structured, locally discriminative cues analogous to RGB. This formulation is extensible: other geometry descriptors (e.g., curvature) may serve as alternative proxies and yield gains. Overall, GCP provides a geometry-driven substitute for the discriminative component of RGB, improving fine-grained geometric perception when color is unreliable.

2.4 Training Strategy

We adopt a curriculum-style two-stage training: Stage-1 learns robust 3D geometry and geometry-language alignment, while Stage-2 improves diagnosis and generation under higher-quality (partially rationale-augmented) supervision.

Stage-1. We train the 3D encoder and projectors and freeze the LLM, leveraging the large-scale Stage-1 data to build strong geometric representations and

Table 2. Comparison with baseline methods on multi-disease IOS diagnosis. The best result is bolded, while the second best result is underlined.

Models	Input	Acc	F1	Preci	Recall	PR
GPT-5 [17]	Multi-View Images	62.26	44.73	49.41	49.20	99.84
Gemini 3 Pro [7]	Multi-View Images	<u>67.65</u>	<u>48.93</u>	**56.78**	<u>51.85</u>	99.71
Qwen3VL-8B [1]	Multi-View Images	57.88	36.46	45.21	46.58	100
InternVL3.5-8B [18]	Multi-View Images	56.23	34.77	40.36	45.47	100
InternVL3.5-14B [18]	Multi-View Images	35.68	24.61	29.14	46.87	100
MedGemma-1.5 [16]	Multi-View Images	52.21	31.31	36.58	45.25	100
HuatuoGPT-V-7B [4]	Multi-View Images	59.26	39.14	40.33	45.56	100
HuluMed-7B [9]	Multi-View Images	61.21	34.68	40.09	45.12	100
HuluMed-14B [9]	Multi-View Images	53.72	36.87	41.63	46.58	100
PointLLM-7B [20]	3D Point Cloud	43.03	34.18	44.62	44.80	99.85
PointLLM-13B [20]	3D Point Cloud	34.57	26.76	42.04	45.03	96.78
ShapeLLM-7B [14]	3D Point Cloud	30.27	19.23	14.66	44.95	100
ShapeLLM-13B [14]	3D Point Cloud	26.13	17.78	13.61	45.10	100
IOSVLM(Ours)	3D Point Cloud	**77.23**	**50.39**	<u>52.19</u>	**52.96**	**100**

stable token alignment. Point features with GCP are used to reduce the gap to color-dependent point-cloud pretraining.

Stage-2. We freeze the 3D encoder and fine-tune the projectors and LLM with LoRA, leveraging higher-quality annotations to further enhance semantic modeling and generative reasoning, while preventing degradation of the learned geometric representations. A unified generative objective is applied: label-only samples supervise y, while rationale samples supervise (y, r).

3 Experimental Results

3.1 Experimental Setup

We evaluate on IOSVQA, with 229,943/15,598/5,884 samples for Stage-1 training, Stage-2 training, and testing, respectively. We report Macro Accuracy (Acc), Macro F1 (F1), Precision (Preci), and Recall, averaged over tasks, and additionally Parsing Rate (PR), measuring the fraction of generated answers that can be parsed into a valid label. IOSVLM uses LLM from Qwen3VL-8B-Instruct [1] with N=10,000 points and 32 learnable visual prompts. LLM from Qwen3VL-8B was selected because it is open-source, reproducible, moderately deployable, instruction-following, and designed for multimodal spatial understanding, suiting geometry-grounded IOS diagnosis.

3.2 Comparison with State-of-the-Arts

Table 2 compares IOSVLM with 4 categories of representative baselines: (i) proprietary multimodal LLMs [17,7], (ii) open-source general-purpose 2D MLLMs

Table 3. Ablation studies. V/P/L denote 3D vision encoder, projector, and LLM, respectively. "∗" indicates tuning with rationales. GCP: Geometry-to-Chromatic Proxy.

Ablation part	Method choice	Acc	F1	Preci	Recall	PR
LLM	Qwen3-4B [21]	71.10	39.22	35.55	45.15	100
	Qwen3-8B [21]	70.73	38.84	36.42	45.26	100
	Qwen3VL-8B	67.02	**43.10**	**48.94**	**46.70**	95.08
GCP	✗	67.02	43.10	48.94	46.70	95.08
	✓	**72.28**	**48.06**	**51.25**	**49.82**	**97.00**
Training mode	Stage 1: V & P	72.28	48.06	51.25	49.82	97.00
	*Stage 2: P & L	**77.23**	**50.39**	**52.19**	**52.96**	**100**
	Stage 2: P & L	77.92	50.35	51.81	53.61	99.74
	Stage 1: V & P & L	75.38	47.79	49.73	51.07	99.97
	*Stage 2: P & L	75.72	50.78	52.37	52.62	100

[1,18], (iii) open-source medical 2D MLLMs [16,4,9], and (iv) open-source general 3D MLLMs operating on point clouds [20,14]. For a fair comparison, all 3D baselines used the same point-cloud preprocessing as IOSVLM, while all 2D baselines were evaluated on conventional multi-view IOS renderings. Following standard orthodontic conventions, occluded-arch scans were rendered into five views: frontal in occlusion, left and right buccal in occlusion, maxillary occlusal, and mandibular occlusal views. Single-arch scans were rendered into four views: frontal, left buccal/lateral, right buccal/lateral, and occlusal views.

Overall performance. IOSVLM achieves the best overall results, obtaining the highest Acc (77.23%) and F1 (50.39%), as well as the best Recall (52.96%). Notably, IOSVLM outperforms all open-source 2D MLLMs by a large margin (at least +16.02% Acc/+11.25% F1) and substantially surpasses open-source 3D MLLMs (at least +34.20% Acc/+16.21% F1), highlighting the advantage of directly modeling native 3D IOS geometry.

Comparison to proprietary MLLMs. Compared with GPT-5, IOSVLM improves by +14.97% Acc and +5.66% F1; compared with Gemini 3 Pro, it yields +9.58% Acc and +1.46% F1 and achieves higher recall (52.96% vs. 51.85%). Notably, IOSVLM uses only an 8B-scale LLM, yet surpasses these substantially larger proprietary MLLMs, highlighting that our gains primarily come from geometry-aware representation and alignment rather than model size. Gemini 3 Pro attains the best precision (56.78%), while IOSVLM provides a more balanced profile with the strongest accuracy-F1-recall, which is crucial for multi-disease diagnosis under class imbalance and fine-grained ambiguity.

Parsing rate (PR). IOSVLM reaches 100% macro PR, indicating consistently parsable outputs in the task-specific label space, while proprietary models show slightly lower PR (99.84%/99.71%). We observe that, despite near-perfect PR, proprietary models still exhibit sporadic empty outputs that break parsability. Together, these results demonstrate that IOSVLM not only improves diagnostic performance but also produces more reliably structured predictions and informative output for large-scale multi-task evaluation.

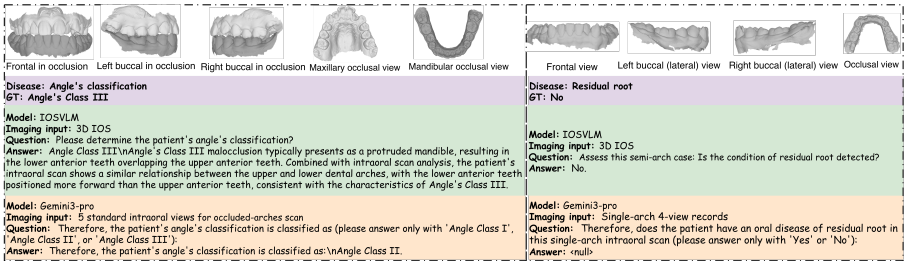


Fig. 3. Qualitative examples of generated diagnostic responses.

3.3 Ablation Studies

Effect of Geometry-to-Chromatic Proxy (GCP). In Table 3, "×" denotes using a constant "white" color (i.e., no separability cue), while "✓" replaces the color channels with the normal-based proxy. GCP yields consistent gains (+5.26% Acc and +4.96% F1), supporting our hypothesis that the discriminative component of RGB, not its semantics, is what benefits pretrained encoders. This also validates normal vectors as an effective instantiation of GCP, and suggests the proxy can be further extended to other geometric cues (e.g., curvature).

Training strategy. Our two-stage curriculum achieves strong performance, with the default setting yielding the best overall accuracy and F1. Interestingly, training V&P&L in Stage-1 outperforms V&P initially, but becomes slightly inferior after Stage-2 in the Acc–F1 trade-off. We attribute this to early LLM updates under mixed-quality supervision, which may bias generation and limit subsequent high-quality adaptation. Nevertheless, all variants are competitive, suggesting that native 3D IOS input with a VLM pipeline forms a strong foundation, while the staged curriculum provides more stable optimization for multi-disease ambiguity.

Rationale supervision: reliability over raw F1. Using rationales in Stage-2 yields comparable macro-F1 to label-only tuning, suggesting that rationales mainly regularize generation rather than improving geometric discriminability. Importantly, rationale tuning improves output reliability/usability: it maintains diagnostic performance while achieving higher macro PR, , and we qualitatively observe fewer degenerate behaviors such as repeated labels. This indicates improved controllability and better-calibrated free-form answers, which is crucial for downstream clinical pipelines.

Choice of LLM. Among LLM candidates, Qwen3VL-8B yields the best macro-F1 under our setting. Although its macro accuracy is slightly lower than the other LLMs, we observe that in our highly imbalanced tasks those LLMs tend to collapse to majority-class predictions, inflating accuracy but hurting minority coverage. In contrast, Qwen3VL-8B exhibits better class-coverage awareness and less majority bias, leading to a more balanced precision-recall profile.

3.4 Qualitative Analysis

In Fig. 3, IOSVLM shows more accurate diagnoses and more reliable, parsable outputs than the strong Gemini 3 Pro, supporting that directly modeling native 3D IOS geometry improves sensitivity to fine-grained morphological cues and enhances output robustness for clinical use.

4 Conclusion

We presented IOSVLM, an end-to-end 3D VLM for unified multi-disease IOS diagnosis and generative VQA, together with IOSVQA, a large-scale benchmark spanning 23 oral diseases and heterogeneous scan types. We introduce a geometry-to-chromatic proxy and a curriculum-style training strategy to better leverage color-dependent point-cloud pretraining and improve robustness. Experiments demonstrate clear gains over strong baselines, supporting direct 3D geometry modeling for practical IOS-based clinical use.

Acknowledgments. This work is supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant no. 2025C01128), and the ZJU-Angelalign R&D Center for Intelligence Healthcare.

Disclosure of Interests. Yang Feng is employed by Angelalign Technology Inc.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Borghi, L., Lumetti, L., Cremonini, F., Rizzo, F., Grana, C., Lombardo, L., Bolelli, F.: Bits2bites: Intra-oral scans occlusal classification. In: Oral and Dental Image aNalysis Workshop (2025)
3. Cai, Z., Zhang, J., Zhao, J., Zeng, Z., Li, Y., Liang, J., Chen, J., Yang, Y., You, J., Deng, S., et al.: Dentalgpt: Incentivizing multimodal complex reasoning in dentistry. arXiv preprint arXiv:2512.11558 (2025)
4. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
5. Eggmann, F., Blatz, M.B.: Recent advances in intraoral scanners. Journal of Dental Research **103**(13), 1349–1357 (2024)
6. Ender, A., Zimmermann, M., Mehl, A.: Accuracy of complete-and partial-arch impressions of actual intraoral scanning systems in vitro. Int J Comput Dent **22**(1), 11–19 (2019)
7. Google DeepMind: Gemini 3 pro model card. Model card (PDF) (Dec 2025), https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf, accessed 2026-02-27

8. Hao, J., Liang, Y., Lin, L., Fan, Y., Zhou, W., Guo, K., Ye, Z., Sun, Y., Zhang, X., Yang, Y., et al.: Oralgpt-omni: A versatile dental multimodal large language model. arXiv preprint arXiv:2511.22055 (2025)
9. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
10. Lian, C., Wang, L., Wu, T.H., Wang, F., Yap, P.T., Ko, C.C., Shen, D.: Deep multi-scale mesh feature learning for automated labeling of raw dental surfaces from 3d intraoral scanners. IEEE transactions on medical imaging **39**(7), 2440–2450 (2020)
11. Liu, Z., He, X., Wang, H., Xiong, H., Zhang, Y., Wang, G., Hao, J., Feng, Y., Zhu, F., Hu, H.: Hierarchical self-supervised learning for 3d tooth segmentation in intra-oral mesh scans. IEEE Transactions on Medical Imaging **42**(2), 467–480 (2023). https://doi.org/10.1109/TMI.2022.3222388
12. Mangano, F., Gandolfi, A., Luongo, G., Logozzo, S.: Intraoral scanners in dentistry: a review of the current literature. BMC oral health **17**(1), 149 (2017)
13. Meng, Z., Hao, J., Dai, X., Feng, Y., Liu, J., Feng, B., Wu, H., Gai, X., Zhu, H., Hu, T., et al.: Dentvlm: A multimodal vision-language model for comprehensive dental diagnosis and enhanced clinical practice. arXiv preprint arXiv:2509.23344 (2025)
14. Qi, Z., Dong, R., Zhang, S., Geng, H., Han, C., Ge, Z., Yi, L., Ma, K.: Shapellm: Universal 3d object understanding for embodied interaction. In: European Conference on Computer Vision. pp. 214–238. Springer (2024)
15. Rodríguez-Ortega, J., Pérez-Hernández, F., Tabik, S.: Charnet: conditioned heatmap regression for robust dental landmark localization. arXiv e-prints pp. arXiv–2501 (2025)
16. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
17. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
18. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
19. Xiong, H., Li, K., Tan, K., Feng, Y., Zhou, J.T., Hao, J., Ying, H., Wu, J., Liu, Z.: Tsegformer: 3d tooth segmentation in intraoral scans with geometry guided transformer. In: International conference on medical image computing and computer-assisted intervention. pp. 421–432. Springer (2023)
20. Xu, R., Wang, X., Wang, T., Chen, Y., Pang, J., Lin, D.: Pointllm: Empowering large language models to understand point clouds. In: European Conference on Computer Vision. pp. 131–147. Springer (2024)
21. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
22. Zanjani, F.G., Moin, D.A., Verheij, B., Claessen, F., Cherici, T., Tan, T., et al.: Deep learning approach to semantic segmentation in 3d point cloud intra-oral scans of teeth. In: International Conference on Medical Imaging with Deep Learning. pp. 557–571. PMLR (2019)
23. Zelikman, E., Wu, Y., Mu, J., Goodman, N.D.: Star: Self-taught reasoner bootstrapping reasoning with reasoning. In: Proc. the 36th International Conference on Neural Information Processing Systems. vol. 1126, pp. 0–55 (2024)

24. Zhang, B., Miao, Y., Wu, T., Chen, T., Jiang, J., Li, Z., Tang, Z., Yu, L., Su, J.: Archmap: Arch-flattening and knowledge-guided vision language model for tooth counting and structured dental understanding. arXiv preprint arXiv:2511.14336 (2025)
25. Zhang, J., Du, B., Miao, Y., Sun, D., Cao, X.: Oralgpt: A two-stage vision-language model for oral mucosal disease diagnosis and description. arXiv preprint arXiv:2510.13911 (2025)